2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Spatial Bias in Vision-Based Voice Activity Detection

Kalin Stefanov
Institute for Creative Technologies
University of Southern California
Los Angeles, USA
kstefanov@ict.usc.edu

Mohammad Adiban* and Giampiero Salvi†
Department of Electronic Systems
Norwegian University of Science and Technology
Trondheim, Norway
mohammad.adiban@ntnu.no* and giampiero.salvi@ntnu.no†

***Abstract*—We develop and evaluate models for automatic vision-based voice activity detection (VAD) in multiparty human-human interactions that are aimed at complementing acoustic VAD methods. We provide evidence that this type of vision-based VAD models are susceptible to spatial bias in the dataset used for their development; the physical settings of the interaction, usually constant throughout data acquisition, determines the distribution of head poses of the participants. Our results show that when the head pose distributions are significantly different in the train and test sets, the performance of the vision-based VAD models drops significantly. This suggests that previously reported results on datasets with a fixed physical configuration may overestimate the generalization capabilities of this type of models. We also propose a number of possible remedies to the spatial bias, including data augmentation, input masking and dynamic features, and provide an in-depth analysis of the visual cues used by the developed vision-based VAD models.**

***Index Terms*—neural networks, vision, voice activity detection, dataset bias, spatial bias**

I. Introduction

Natural and effective face-to-face human-human interactions require smooth coordination of the changes in the roles on the conversational floor (*i.e.*, speaker, addressee, bystander), known as footing [1], [2]. Since clear conversational roles in face-to-face communication are vital for smooth and effective interaction, a machine which is aware of the established roles in real-time could avoid misunderstandings or talking over other participants. Therefore, machines need the ability to perform accurate voice activity detection in conversations with overlapping speech and multiple parties. Furthermore, such voice activity detection abilities should minimize any assumptions for the environment in which the machine will engage in an interaction. Such assumptions include the level of noise, number of participants, and spatial configuration (placement of participants and sensors).

In this paper, we develop and analyze models for voice activity detection based solely on visual input (*i.e.*, RGB face data) using state-of-the-art Convolutional Neural Networks.

This work was supported in part by the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

We would like to acknowledge the NVIDIA Corporation for donating some of the GeForce GTX TITAN GPUs used for this research.

Giampiero Salvi is also affiliated with School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden.

Fig. 1. Example of a vision-based voice activity detector. The first and second images capture negative head rotations around the Y-axis (yaw) w.r.t. the camera origin, while the third and fourth images represent positive rotations.

Since multiparty face-to-face interactions are situated in space, we argue that this type of voice activity detectors incorporate information about the spatial bias captured in the data used for training the detectors, that is, information about the spatial configuration or seating arrangement might bias the classifiers. In other words, the performance of the detectors on data sampled from significantly different spatial distribution (or seating arrangement) will decrease significantly, thus the detectors will have poor generalization capabilities. This work provides evidence for this phenomenon and to the best of our knowledge is the first to identify this bias and investigate different mitigation strategies.

The main contributions of this work are:

- We demonstrate that spatial bias (*i.e.*, head pose information) related to the physical settings of the interaction is encoded in the learned representations of certain types of vision-based voice activity detectors.
- We analyse the effect of data augmentation, input masking and dynamical inputs on the generalization capabilities of the models with mismatched train and test data.
- We perform in-depth analysis of the features extracted by the models in order to explain our experimental results.

The rest of the paper is organized as follows. First, in Section II, we outline previous research on voice activity detection and publicly available datasets for building vision-based voice activity detectors. We describe the developed models in Section III. The experimental setup and evaluation are given in Section IV and the results of the conducted experiments are presented in Section V. Discussion on the limitations and contributions of the developed models is given in Section VI. We conclude the paper in Section VII.

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9413345

II. Related Work

In this section, we first discuss existing work on automatic voice activity detection. We then turn our focus on the few publicly available datasets suitable for developing vision-based voice activity detectors.

A. Voice Activity Detection

Voice activity detection (VAD) is the task of determining if a certain speaker is active at any point in time. If multiple speakers are present, the more complex task of determining which speaker is active is called diarization. VAD is important for many applications and in each area, different constraints are imposed on the methods. Previous work includes audio-only, video-only and audio-visual approaches to VAD.

In clean acoustic conditions, and with single speaker inputs, the acoustic information is fundamental for the VAD task, and methods for audio-only VAD have been extensively studied. Anguera *et al*. [3] and Tranter and Reynolds [4] offer comprehensive reviews of the research in this field. Audio-only VAD systems usually suffer from noisy environments, far field microphones (as in meeting settings) and by speakers that overlap in time. Additionally, audio-only approaches are more limited in multiparty interactions, where it is important to assign the detection to speakers that might be close physically.

In order to address the shortcomings of audio-only VAD, many video-only and audio-visual approaches have been proposed in the literature. Video-only methods attempt to directly model the face [5], [6] or some aspects of the face (*e.g*., lip movements [7]) in order to detect the voice activity. The drawbacks of these types of methods are related to a number of motions including facial expressions, yawning or chewing that can be misinterpreted as speaking.

Audio-visual voice activity detection combines information from both the audio and the video signals. The idea is that by complementing the audio approach with its video counterpart, the performance will be generally better because of increased robustness [8]–[12]. The application of audio-visual synchronization to speaker detection in broadcast videos was explored by Nock *et al*. [13]. Unsupervised audio-visual detection of the speaker in meetings was proposed in Friedland *et al*. [14]. Zhang *et al*. [15] presented a boosting-based multimodal speaker detection algorithm applied to distributed meetings. Mutual correlations to associate an audio source with regions in the video signal were demonstrated by Fisher *et al*. [16], and Slaney and Covell [17] showed that audio-visual correlation can be used to find the temporal synchronization between the audio signal and the speaking face. An elegant solution was proposed in Hershey and Movellan [18] where the mutual information between the acoustic and visual signals is computed by means of a joint multivariate Gaussian process, with the assumption that only one audio and one video streams were present and that locating the source corresponds to finding the pixels in the image that correlate with acoustic activity.

In more recent studies, researchers have employed artificial neural network architectures to build voice activity detectors from audio-visual input. A multimodal Long Short-Term Memory (LSTM) model that learns shared weights between modalities was proposed in Ren *et al*. [12]. The model was applied to speaker naming in TV shows. A combination of pre-trained Convolutional Neural Network (CNN) model used for the image encoder and an LSTM model used for the classifier was presented in Stefanov *et al*. [19]. Stefanov *et al*. [20] further proposed a self-supervised method for vision-based voice activity detection in the context of language acquisition. Hu *et al*. [21] proposed a CNN model that learns the fusion function of face and audio information. Roth *et al*. [22] introduced a new audio-visual dataset for voice activity detection and Chung [23] proposed a method for active speaker detection on that dataset.

Other approaches to voice activity detection include a general pattern recognition framework used by Besson and Kunt [24] applied to the detection of the speaker in audio-visual sequences. Visual activity (the amount of movement) and the focus of visual attention were used as inputs by Hung and Ba [25] to determine the current speaker on real meetings. Stefanov *et al*. [6] used facial action units as inputs to hidden Markov models to determine the active speaker in multiparty interactions and Vajaria *et al*. [26] demonstrated that information from body movements can improve the detection performance.

B. Publicly Available Datasets

Recently several datasets have been created and made publicly available in order to build and evaluate audio-visual voice activity detectors. The AVSpeech dataset [27] is an automatically collected large-scale dataset consisting of several lecture recordings. The recordings capture around 4700 hours of audio-visual data with a single clearly visible face and the corresponding audio. However, in order to be used for development and evaluation of vision-based voice activity detectors, the data needs to be labeled.

The Columbia dataset [11] consists of a recording of a panel discussion between seven speakers labeled with speaking and not speaking state. The drawback of this dataset is its small size that makes it less useful for developing machine learning models.

The AVA-ActiveSpeaker dataset [22] consists of around 38 hours of audio-visual data. Each of the 3.65 million frames is manually labeled for speaking and audible, speaking but not audible, and not speaking state. The drawback of this dataset is the fact that the data is noisy and in some cases the audio and video signals are out of sync.

The AMI dataset [28] consists of 100 hours of meeting recordings. It is manually annotated for many different phenomena, including orthographic transcriptions, hence accurate voice activity labels could be generated automatically.

In this study we use an in-house dataset containing three-way interactions collected through three cameras. Further details on the dataset are given in Section IV-A.

III. Methods

The goal of the methods described in this section is to detect the speaking state (*i.e.*, speaking or not speaking) of all visible faces in a multiparty interaction, using only visual information (*i.e.*, the RGB data).

A. Problem Definition

Given a number of speakers, and a number of sensors (cameras), vision-based voice activity detection consists of the task of determining at any point in time, which speakers are active from the video streams. If each camera only captures a single speaker, or face tracking is available, this is a binary classification problem where the input is the part of the image assigned to each speaker. An example output of a vision-based voice activity detector is illustrated in Figure 1.

B. Models

We formulate the problem of detecting the state of a face image (*i.e.*, speaking or not speaking) as a binary classification task. The models developed in this study consist of a state-of-the-art Convolutional Neural Network serving as an image encoder followed by two fully-connected layers that classify the obtained image representations into one of the two classes. For learning the representations of the input face images we use a truncated ResNet-18 [29] as a basis for the encoder (by removing the classifier from the original model). For training the models, first, we initialize the truncated ResNet-18 model with the pre-trained weights on ImageNet [30] and then simultaneously fine-tune the encoder and train the classifier with our data.

IV. Experiments

In this section, we first describe the dataset used to train and evaluate the voice activity detectors. We then provide the general setup of the conducted experiments.

A. Dataset

The models are trained and evaluated with a multimodal multiparty dataset, described in [31]. The spatial configuration of the recordings is shown in Figure 2. Three participants take part in each session, where two of the participants interact with a moderator (the third participant). Each participant was recorded by a camera positioned in front of him/her. A total of 15 sessions were recorded, each with a duration of ∼30 minutes, resulting in ∼7.5 hours of data per recording device. The moderator is the same for all sessions, whereas the other participants vary for a total of 24 unique participants. The interactions occur around a round table and the participants are seated. The spatial configuration is constant throughout the data collection. All interactions are in English and all data streams are spatially and temporally synchronized and aligned. The dataset is augmented with information about the head rotation around the Y-axis (yaw) obtained with OpenFace [32]. The polar plots in Figure 2 show the distribution of head rotations over all recording sessions and for each participant location, which is relevant for the rest of this study.

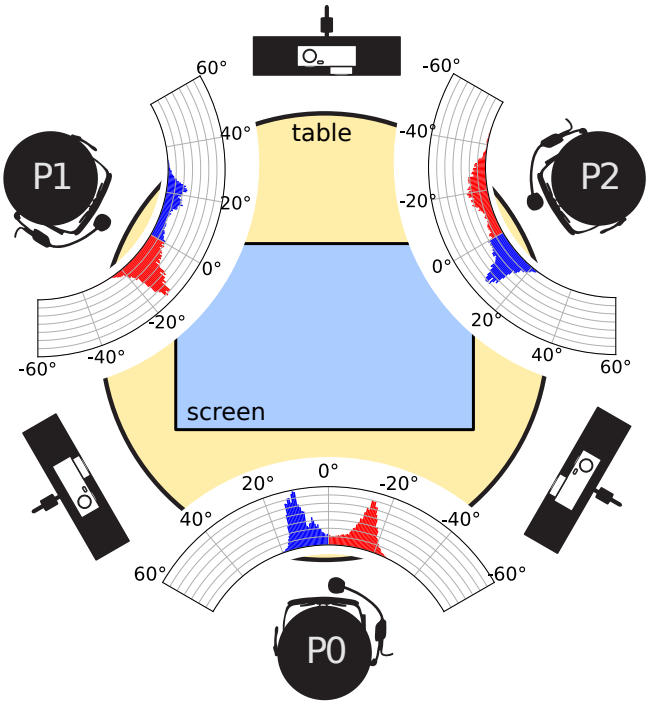

Fig. 2. Spatial configuration of the sensors and participants in the dataset. Each spoken interaction involves three participants seated at the positions denoted with P0, P1, and P2. The polar plots show the distribution of head rotations around the Y-axis (yaw) for each position, estimated over all sessions in the dataset. Angles are relative to the direction of the corresponding camera. Positive angles are plotted in blue and negative angles in red.

Here we consider the RGB video stream generated by the Kinect v2 device pointed at each participant and the audio stream generated by the participant's close-talking microphone. The voice activity labels are obtained by manual annotations of the audio streams.

B. Setup

The experiments are designed to test the models in the spatial position P0 that corresponds to the moderator of the spoken interactions, see Figure 2. This provides the opportunity to produce well trained speaker-dependent models and test the models in details. The total number of frames used in the experiments is 411,356 for a total duration of analyzed video data of approximately 4 hours at 30 frames per second.

In the experiments we keep the model architecture constant and vary the inputs with which each model is trained and tested, in order to give insights into the model's capabilities and limitations.

The first two variables in the experiments, zoom level and input kind are illustrated in Figure 3 and determine the visual representation given as input to the models. The zoom level is obtained by masking potentially irrelevant information from the original images and results in four alternative representations: FRAMES, ALIGNS, FACES, and LIPS (columns in the figure). The input kind is static when we use the original images (first row in the figure), or dynamic as in the second

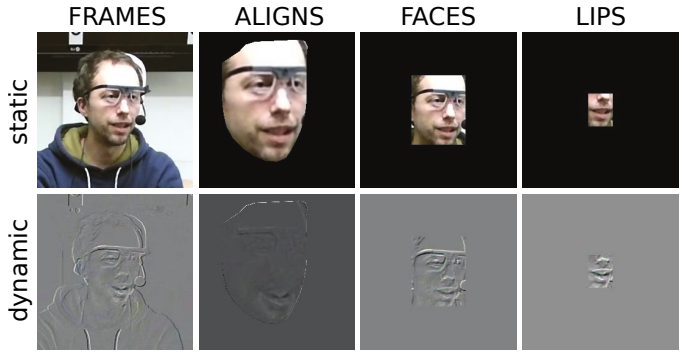

Fig. 3. Different types of input data considered in this study. The first row illustrates the static inputs and the second row the dynamic inputs. Different columns show different levels of zoom (or masking): starting from left, these are denoted in the paper by FRAMES, ALIGNS, FACES, and LIPS.

row. The dynamic inputs are generated using dynamic image networks [33] to summarize a sequence of frames into one image. We slide a fixed window (6 frames, ∼200ms) with a hop of one frame over the static inputs. This results in the same amount of input data in both the static and dynamic cases.

Another variable used in our experiments is the partitioning of the train and test data introduced with the aim to generate mismatch between the train and test data with respect to head rotations. We split the train and test data into POS and NEG partitions depending on the positive or negative sign of the head rotation angles (see Figure 2 for an illustration). The full train and test sets are referred to with FULL in the text and figures.

Finally, in an attempt to mitigate the effects of mismatch between train and test data, we perform a simple procedure for data augmentation where we create a new train set from a POS or NEG partition by flipping all the images horizontally. This training is referred to in the following as AUGMENTED whereas the original partitions are labeled ORIGINAL.

For each experiment, we use a 10-fold cross-validation procedure to randomly (but preserving the underlying distribution of head rotations) split the data into three parts: train, validation, and test data. The proportions of the three partitions are respectively ∼80%, ∼5%, and ∼15%. For training the models we used Adam [34] optimizer with default parameters ($\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$) and cross-entropy loss function. Each model is trained for 25 epochs when using the original dataset and 50 epochs in case of the augmented dataset (it is exactly two times larger than the original dataset). The models corresponding to the best validation performance are selected for evaluation on the test set. All models are implemented in PyTorch [35].

Each model outputs a posterior probability over the two possible outcomes (*i.e.*, speaking or not speaking). Since the goal is a binary classification, the detection of the positive voice activity happens when the corresponding probability exceeds 0.5. The evaluation of each model is performed by computing the F1 score on a frame-by-frame basis.

V. Results

In this section we report the experimental results. We first discuss the results obtained by training the models on the full train set. In Figure 4 the F1 scores obtained over 10-fold cross-validation are presented with box plots in different experimental conditions. The left plot shows models trained with static inputs, whereas the right plot with dynamic inputs. The X-axis shows the level of details in the input images, from the broader FRAMES to the more detailed LIPS (see Figure 3 for a visual representation). Finally, the different test sets are color coded and correspond to the full test set (FULL) and the partitions of the test set containing negative (NEG) or positive (POS) head rotations.

From the figure we can make a number of general observations: i) results with dynamic inputs are in general worse than those with static inputs, ii) focusing on more details of the face or the lips has a negative impact on the results compared to using the full images, iii) results are very stable across repetitions (concentrated around the median), iv) for the static inputs results do not change if we restrict the test set to only negative or positive rotations. However, for the dynamic inputs, the negative rotations consistently obtain slightly better results, and v) results with FRAMES and ALINGS static inputs have the highest performance and approach 98% F1 scores.

In order to verify if this very high performance might be affected by the spatial bias in the dataset as discussed in Section I, we present results obtained by splitting the train set as well as the test set into negative and positive angles of rotations. Figure 5 shows the F1 score when the train and test sets are either matched (both subsets with negative or both with positive angles) or mismatched. Again, all results are presented with box plots over 10-fold cross-validation. The left and right plots refer to static and dynamic inputs respectively. The X-axis corresponds to the level of zoom in the inputs. Finally the color code corresponds to the train and test set selection. In the ORIGINAL condition, either only examples of negative rotations or positive rotations are included in the train set. In the AUGMENTED condition the same train sets of the ORIGINAL conditions are augmented by flipping the images horizontally. In the matched conditions, both the train and test set contain examples with the same direction of rotation. In the mismatched condition the rotations in the test set are in the opposite direction of those in the train set.

The matched conditions present a similar pattern seen in Figure 4, in spite of the fact that the amount of train data is around half in the ORIGINAL case. However, there is a consistent and considerable drop in performance (up to 20 percentage points) when the test set is mismatched with the train set. Data augmentation consistently improves the situation in most conditions. However, it comes short of solving the problem. Surprisingly, this drop in performance is observed even when concentrating on details of the face (ALIGNS, FACES and LIPS conditions). Using dynamic inputs does not improve the situation as performance is lower in general.

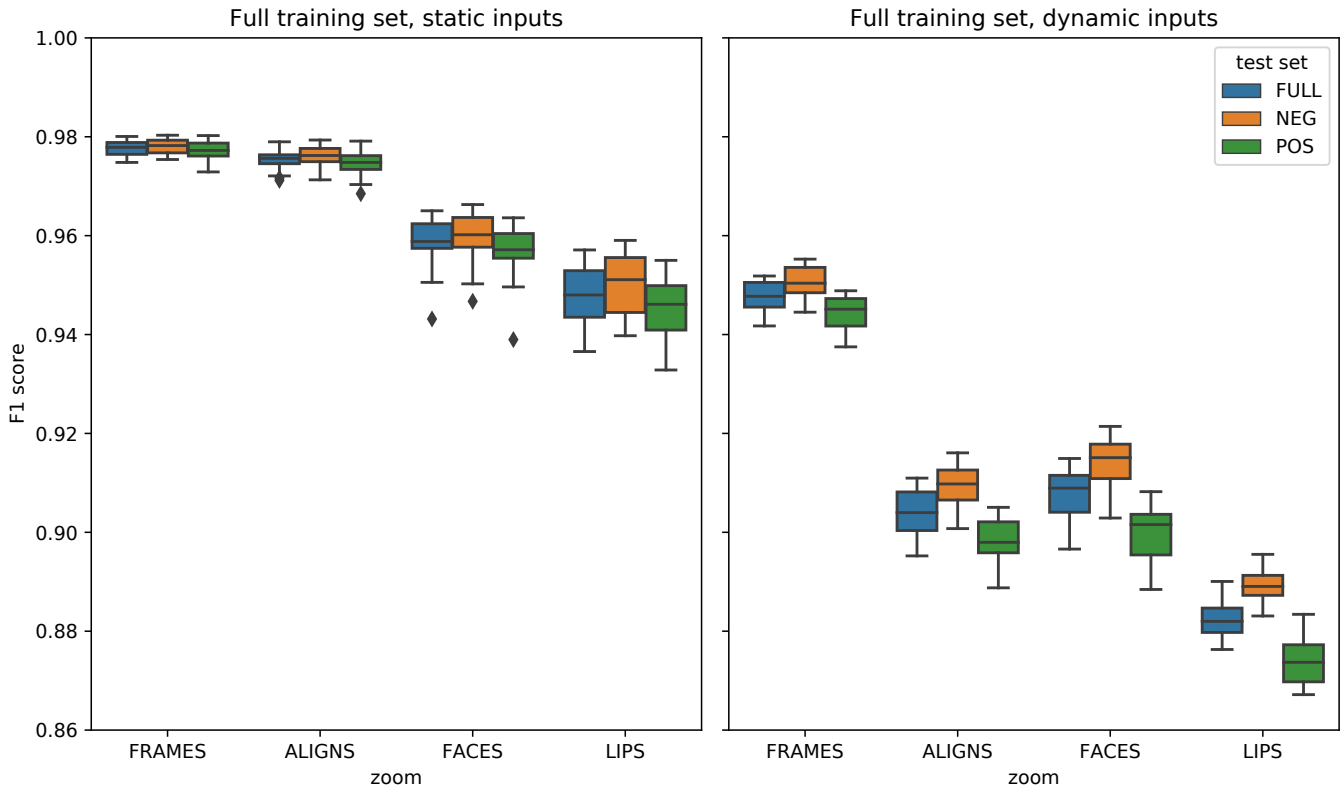

Fig. 4. Box plots of the F1 score over 10-fold cross-validation as a function of zoom level (FRAMES, ALIGNS, FACES, LIPS) and test set (FULL, POS, NEG). The models are trained on the full train set with static (left) or dynamic inputs (right).

In the case of ORIGINAL static inputs we also report in Figure 6 a detailed version of the results where we have split the matched and mismatched conditions into the four combinations of train and test partitions. The figure shows that training on positive or negative angles has little influence and the main factor is if the test set is matched or mismatched to the train data.

VI. Discussion

The results in Figure 4 on the full train and test set with the full image input (FRAMES) may suggest that the models learn the visual voice activity detection task with very high level of performance and generalize well between the train and test set. However, this global evaluation turns out to be misleading as one digs further into the details. The results shown in Figure 5 for mismatched train and test sets demonstrate how it is sufficient to change the head rotation angle to confuse the models if those angles were not observed during training. This suggests that spatial bias (*i.e.*, head pose information) related to the physical settings of the recordings is encoded into the learned representations of this type of vision-based voice activity detectors. These representations fail to distill the information that is relevant to the VAD task, disregarding other aspects of the input images even in spite of the fact that all other variables (illumination, subject, experimental setting), are constant during our experiments.

We proposed a number of approaches to mitigate this problem showing that this is a hard problem to solve.

Data augmentation by flipping the images horizontally has a positive influence. However, the improvements in performance are limited, and may be overestimated because of the symmetry of the head rotation distribution for position P0 as illustrated in Figure 2. This means that flipping the images creates a distribution of head poses that is similar to that of the full train set. The aspect of the flipped face may still be very different from the original rotated face because of asymmetries in the human face.

Another attempt is to mask parts of the image that should not be involved in the task (ALIGNS, FACES and LIPS conditions). The results presented in Figure 5 demonstrate that this is counter-productive, and suggest that the model might be using information from the input images outside the area of the speaker's face. In order to verify this, we have created saliency maps showing the areas of the image that the model deems more salient to solve the task. Figure 7 shows an example sequence using FRAMES inputs where the speaker is inactive in the first row and active in the second. As expected, the model uses information from the whole picture (including the background and torso of the speaker) and not only the face. This is probably due to the fact that speaker movements are correlated with voice activity and provide useful information for the task. Explaining why pixels

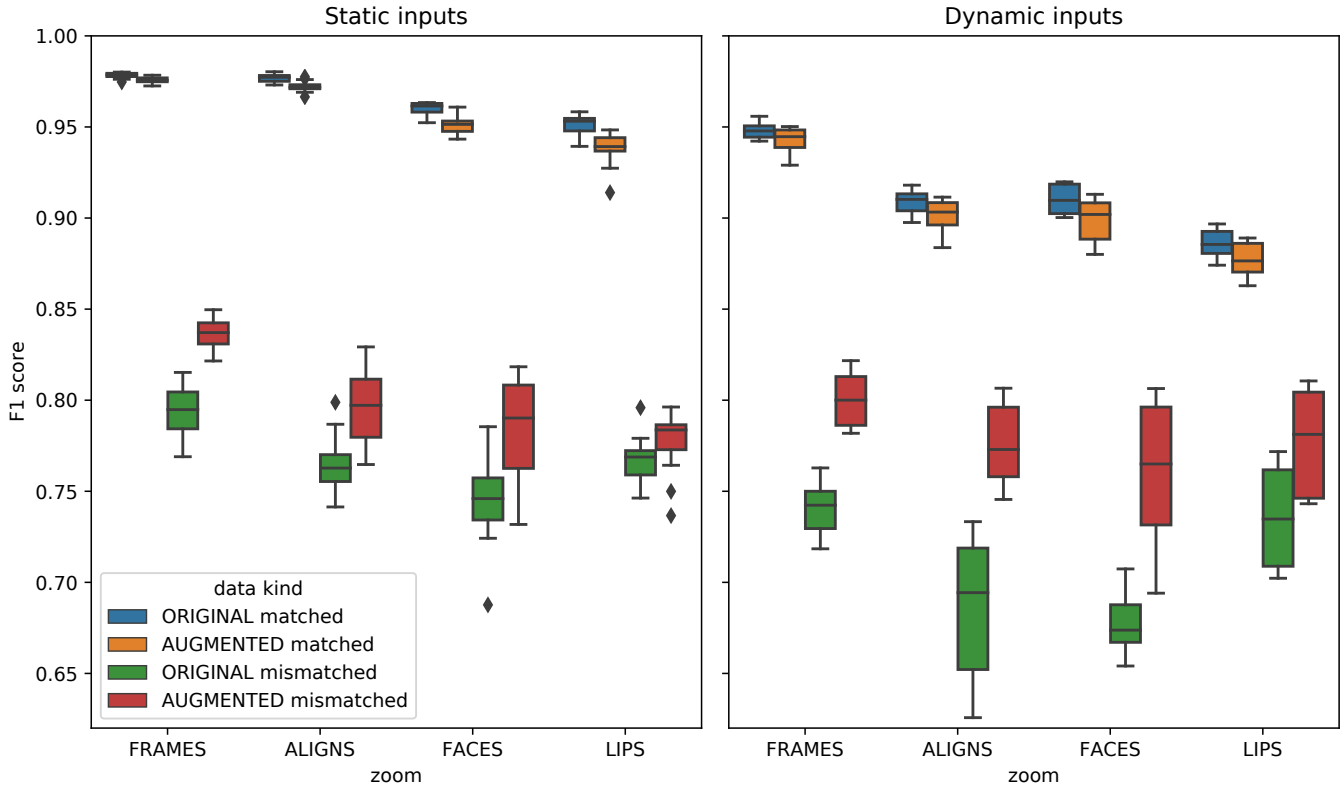

Fig. 5. Box plots of the F1 score over 10-fold cross-validation with matched and mismatched train and test sets. The left plot shows results obtained with static video frames as input to the models, on the right dynamic frames are used. In each plot, the X-axis shows the zoom levels (FRAMES, ALIGNS, FACES, LIPS). The color codes distinguish data kinds. The ORIGINAL data contains examples with either only negative or positive yaw angles both in the train and test set. The AUGMENTED data contains those examples and the ones obtained by flipping the images horizontally. Matched results have the same direction of head rotation in the train and test set, mismatched results have opposite rotations.

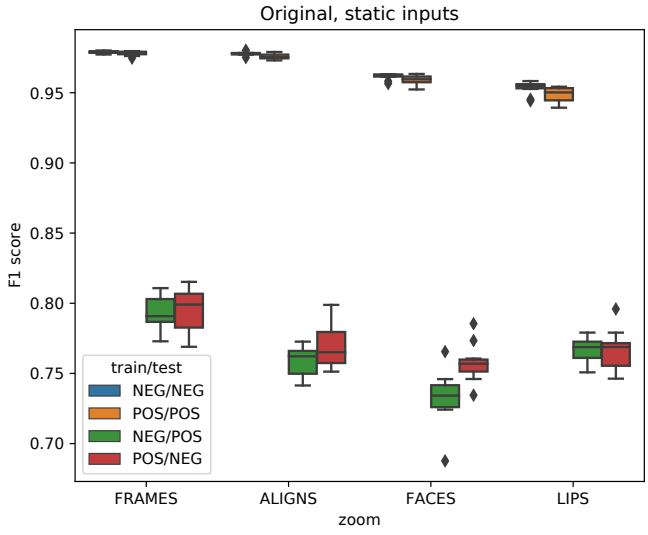

Fig. 6. Box plots of the F1 score over 10-fold cross-validation of the detailed results for ORIGINAL static inputs. All four combinations of train and test partitions are shown. NEG/NEG and POS/POS correspond to the matched condition in Figure 5, whereas NEG/POS and POS/NEG correspond to the mismatched condition.

from the background are also salient is more difficult. One hypothesis is that the speaker movements cause the camera to adjust exposure parameters which creates visual artifacts in the background that are correlated with those movements.

Finally the use of dynamic inputs did not improve the results. One hypothesis is that the algorithm removes important information contained in the static inputs as illustrated in Figure 3. A possible solution would be to stack the static and dynamic inputs into a larger input vector, but this has not been tried so far. We also suspect that there might be artifacts introduced by the algorithm that extracts such images over a number of consecutive frames. However, a conclusive explanation would require a more in-depth analysis.

Since data collection, similar to what we used in this study, is a costly process and requires the use of human resources, data augmentation methods are logical directions in future work. Given the form of the problem we face in this study, in addition to the aforementioned methods, we intend to use homography [39] (or projective transformation) techniques i) for the purpose of data augmentation and ii) to increase the performance of the vision-based voice activity detectors. An example of a homography technique applied to images extracted from the dataset is shown in Figure 8.

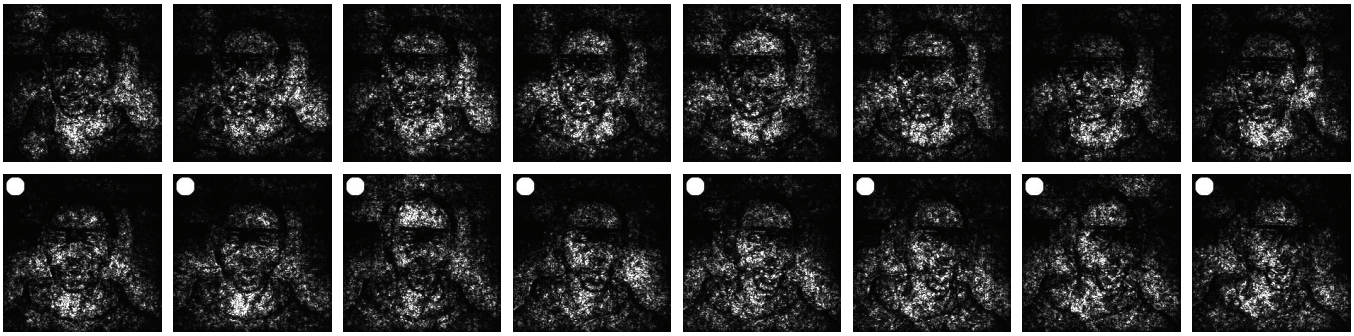

Fig. 7. Saliency maps for a sequence of static FRAMES inputs. The white circle indicates an active speaker. The images are generated by using the methods described in [36] and [37]. The implementation is based on [38]. The result shows that the models use information from the whole picture (including the background and torso of the speaker) and not only the face. This is probably due to the fact that speaker movements are correlated with the voice activity.

Fig. 8. Example of estimating a 2D homography (or projective transformation) from a pair of images. The first row illustrates the static input frames to the 2D homography method. The second row illustrates the rotated and reconstructed versions of the images in the first row using a 2D homography technique. Similar techniques could be used for the purpose of data augmentation in order to increase the performance of vision-based voice activity detection methods.

In order to provide further support for the negative effects of dataset spatial bias on this type of vision-based voice activity detection methods, we intend to perform similar analysis on the rest of the participants present in the used dataset, also minimizing the possible effects that the appearance and behavior of a specific person might have on the models (account for person-specific bias).

Additionally it would be interesting to see to what extent similar spatial bias might impact other datasets and methods for video-only and audio-visual voice activity detection proposed in the literature.

VII. Conclusions

Voice activity detection is a fundamental prerequisite for any machine-based conversational system. Automatic vision-based VAD (based on visual information from the face) in multiparty human-human interactions could complement an acoustic VAD method, thus improving the system robustness in noisy conditions and allowing it to detect an arbitrary number of possibly overlapping active speakers.

In this paper we analyse the problem of spatial bias in vision-based voice activity detection for multiparty spoken interactions. The head pose distribution is an inherent bias in multiparty interaction data, due to the fixed seating arrangement and camera placement. We claim that the performance evaluations observed in our results, as well as those reported in the literature on similar datasets, may be boosted by the similarity of the distribution of the head pose in the train and test sets. By artificially creating a mismatch between the head pose distribution in the train and test data, we show how the performance of a vision-based voice activity detector can be drastically reduced. Although splitting the data between positive and negative head rotations may seem artificial, the problem of spatial bias in datasets with fixed seating arrangements should be evident when looking at the peaked distributions of head rotations in Figure 2.

We propose a number of approaches to mitigate this problem. Our simple data augmentation method provides a consistent but limited improvement. We suggest that more advanced augmentation methods may improve the results even further.

We also show how masking potentially irrelevant information outside the face of the speaker does not improve the results for this kind of modelling. In order to explain this, we analyse the model activations in detail. We show that the models are able to make use of information that is spread in several regions of the input image, and we provide a possible interpretation.

REFERENCES

[1] E. Goffman, *Frame analysis: An essay on the organization of experience.* Harvard University Press, 1974.

[2] ——, *Forms of talk*. University of Pennsylvania Press, 1981.

[3] X. Anguera, N. Bozonnet, S.and Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: a review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[4] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[5] R. Ahmad, S. P. Raza, and H. Malik, "Visual speech detection using an unsupervised learning framework," in *Proceedings of the International Conference on Machine Learning and Applications*, vol. 2, 2013, pp. 525–528.

[6] K. Stefanov, A. Sugimoto, and J. Beskow, "Look who's talking: Visual identification of the active speaker in multi-party human-robot interaction," in *Proceedings of the Advancements in Social Signal Processing for Multimodal Interaction*, 2016, pp. 22–27.

[7] S. Siatras, N. Nikolaidis, M. Krinidis, and I. Pitas, "Visual lip activity detection and speaker detection using mouth region intensities," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 1, pp. 133–137, 2009.

[8] V. P. Minotto, C. R. Jung, and B. Lee, "Simultaneous-speaker voice activity detection and localization using mid-fusion of svm and hmms," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1032–1044, 2014.

[9] R. Cutler and L. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, vol. 3, 2000, pp. 1589–1592.

[10] P. Chakravarty, S. Mirzaei, T. Tuytelaars, and H. Van hamme, "Who's speaking? audio-supervised classification of active speakers in video," in *Proceedings of the ACM on International Conference on Multimodal Interaction*, 2015, pp. 87–90.

[11] P. Chakravarty and T. Tuytelaars, "Cross-modal supervision for learning active speaker detection in video," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 285–301.

[12] J. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan, "Look, listen and learn - a multimodal LSTM for speaker identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016, pp. 3581–3587.

[13] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: an empirical study," in *Proceedings of the International Conference on Image and Video Retrieval*, 2003, pp. 488–499.

[14] G. Friedland, C. Yeo, and H. Hung, "Visual speaker localization aided by acoustic models," in *Proceedings of the ACM International Conference on Multimedia*, 2009, pp. 195–202.

[15] C. Zhang, P. Yin, Y. Rui, R. Cutler, P. Viola, X. Sun, N. Pinto, and Z. Zhang, "Boosting-based multimodal speaker detection for distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 8, pp. 1541–1552, 2008.

[16] J. W. Fisher, T. Darrell, W. T. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Advances in Neural Information Processing Systems 13*, 2001, pp. 772–778.

[17] M. Slaney and M. Covell, "FaceSync: a linear operator for measuring synchronization of video facial images and audio tracks," in *Advances in Neural Information Processing Systems 13*, 2001, pp. 814–820.

[18] J. Hershey and J. Movellan, "Audio-vision: using audio-visual synchrony to locate sounds," in *Advances in Neural Information Processing Systems*, 2000, pp. 813–819.

[19] K. Stefanov, J. Beskow, and G. Salvi, "Vision-based active speaker detection in multiparty interaction," in *Proceedings of the Grounding Language Understanding*, 2017, pp. 47–51.

[20] ——, "Self-supervised vision-based detection of the active speaker as support for socially-aware language acquisition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 2, pp. 250–259, 2020.

[21] Y. Hu, J. S. Ren, J. Dai, C. Yuan, L. Xu, and W. Wang, "Deep multimodal speaker naming," in *Proceedings of the ACM International Conference on Multimedia*, 2015, pp. 1107–1110.

[22] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi, and C. Pantofaru, "Ava-activespeaker: An audio-visual dataset for active speaker detection," *arXiv preprint arXiv:1901.01342*, 2019.

[23] J. S. Chung, "Naver at activitynet challenge 2019 – task b active speaker detection (ava)," *arXiv preprint arXiv:1906.10555*, 2019.

[24] P. Besson and M. Kunt, "Hypothesis testing for evaluating a multimodal pattern recognition framework applied to speaker detection," *Journal of NeuroEngineering and Rehabilitation*, vol. 5, no. 1, p. 11, 2008.

[25] H. Hung and S. O. Ba, "Speech/non-speech detection in meetings from automatically extracted low resolution visual features," Idiap, Tech. Rep., 2009.

[26] H. Vajaria, S. Sarkar, and R. Kasturi, "Exploring co-occurrence between speech and body movement for audio-guided video localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1608–1617, 2008.

[27] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, 2018.

[28] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus: A pre-announcement," in *Proceedings of the Machine Learning for Multimodal Interaction*, 2006, pp. 28–39.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[31] K. Stefanov and J. Beskow, "A multi-party multi-modal dataset for focus of visual attention in human-human and human-robot interaction," in *Proceedings of the International Conference on Language Resources and Evaluation*, 2016.

[32] T. Baltrusaitis, P. Robinson, and L. P. Morency, "Openface: An open source facial behavior analysis toolkit," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2016, pp. 1–10.

[33] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3034–3042.

[34] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," *Computing Research Repository*, vol. abs/1412.6980, 2014.

[35] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proceedings of the NeurIPS Autodiff Workshop*, 2017.

[36] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[37] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," *arXiv preprint arXiv:1605.01713*, 2016.

[38] U. Ozbulak, "Pytorch cnn visualizations," https://github.com/utkuozbulak/pytorch-cnn-visualizations, 2019.

[39] E. Dubrofsky, "Homography estimation," *Diplomová práce. Vancouver: Univerzita Britské Kolumbie*, vol. 5, 2009.